Volume 2, Issue 11, November 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Acoustic Source Localisation System using Microphone Arrays

Ashok Kumar[1] and Maninder Pal[2]
[1] Research Fellow, Ministry of Earth Sciences, New Delhi, INDIA
[2] Department of Electronics & Communication Engineering,
Maharishi Markandeshwar University, Mullana (Ambala), INDIA

Abstract: ***This paper focuses on developing an acoustic source localisation system using commonly available sound monitoring and recording devices such as microphones and sound cards. More specifically, this paper is focused on locating source producing speech signals. For this purpose, the commonly available devices such as microphones and piezo-sensors are used as sensors and single channel USB sound cards are used for digitising analog signals obtained from sensors. The signals obtained from sensors are pre-amplified prior to digitization. The time delay estimation (TDE) using correlation technique is used for estimating the time delay of speech signals arrived at each of the sensors. The localization parameters of the sound source such as azimuth and distance can be calculated from the TDE information. The idea is that if the source is at different distance location from the monitoring sensors, then the signals produced by this source takes different time to reach to sensors and using TDE and multi-sensors, the location of the source can be easily identified. It has also been noticed that the speech signals gets easily distorted by the ambient noise such as noise of a fan, running machinery or vehicles. These noises may significantly affect the signal-to-noise (SNR) and thus may even give incorrect correlation results required for time delay estimation. In is because the mentioned noises closely resemble the speech signals, and produced by a single source. For this purpose, a minimum mean-square error (MMSE) estimator is also proposed in this paper. Thus in summary, a microphone array configured with sensors and usb sound cards is investigated theoretically and experimentally to evaluate its performance in locating sound sources.***

Keywords: Acoustic noise, MMSE estimator, source localisation, microphone arrays, time delay estimation.

I. INTRODUCTION

Speech, in general, can be defined as a mechanism of expressing thoughts and ideas using vocal sounds [1,2]. In humans, speech sounds are produced when breath is exhaled from the lungs & causes either a vibration of the vocal cords (when speaking vowels) or restriction in the vocal tract (when speaking consonants) [3,4]. In general, speech production and perception is a complex phenomenon and uses many organs such as lungs, mouth, nose, ears & their associated controlling muscles and brain. The bandwidth of speech signals is roughly around 4 KHz. However, the human ear can perceive sounds, with frequencies in between 20 Hz to 20 KHz. The signals with frequencies below 20 Hz are called subsonic or infrasonic sounds, and above 20 KHz are called ultrasonic sounds. This paper focuses on locating a source of speech signals using microphone arrays. The noise produced by various sources such as vehicles also lies in the frequency range of speech signals. Therefore, speech signals get easily distorted by the ambient noise or AWGN. These distorted or degraded speech signals are called noisy speech signals. The time delay estimation using correlation, in presence of noise, is usually very difficult and can give incorrect results. Therefore, this paper focuses on speech enhancement of the noisy speech signals, so that the speech source can be easily located even in presence of noise. Some of the key speech enhancement techniques include spectral subtraction approach [5,6,7], signal subspace approach [8,9], adaptive noise cancelling and iterative Wiener filter. The performances of these techniques depend on the quality and intelligibility of the processed final speech signal, and the same has been reported in many articles and journals. The prime focus of all these techniques is to improve signal-to-noise ratio of speech signals. However, these techniques are reported to have several drawbacks such as residual noise and musical sounds. Therefore, this paper uses a MMSE estimator to improve the SNR of the estimated speech signal. This is further discussed in this paper.

II. SYSTEM MODEL

Microphone array consists of a set of microphones to record speech signals from various directions. It is assumed that the as microphones are positioned in different directions, so the position of source will be different with respect to each microphone. Thus, the speech signals will take different time of arrival at each microphone. This difference in time of arrival can be estimated using correlation techniques. Using this time difference, the position of speech source relative to one microphone can be identified. However, noise can significantly affect the time delay estimation using correlation. Therefore,

this paper is divided into two parts: 1) First part reduces the noise from the speech signals recorded through microphones, and 2) second part uses the time delay between the improved quality speech signals to locate the speech source. It is to be noted that the source location is only done in a two dimensional plane; and the azimuth and elevation is not estimated in this paper.

A. Step 1: Improving the SNR using the Laplacian Based MMSE Estimator

To estimate and improve the quality of the speech signals, a MMSE algorithm is used, as described below. In order to derive the MMSE STSA estimator, the *a priori* probability distribution of the speech and noise Fourier expansion coefficients are assumed, as these are unknown in practice. Let y(n) = x(n)+d(n) be the sampled noisy speech signal consisting of the clean signal x(n) and the noise signal d(n). Taking the short-time Fourier transform of y(n), to have:

$$Y(\omega_k) = X(\omega_k) + D(\omega_k) \tag{1}$$

Where, $\omega_k = 2\pi k/N$, k=0,1,2,…N-1, and N is the frame length. The above equation can also be expressed in polar form as

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \tag{2}$$

As, the spectral components are assumed to be statistically independent, the MMSE amplitude estimator $\hat{X}_k$ can be derived from $Y(\omega_k)$ only. That is,

$$\hat{X}_k = E\{X_k|Y(\omega_0), Y(\omega_1), \dots\} \tag{3}$$

$$= E\{X_k|Y(\omega_k)\} = \frac{\int_0^\infty \int_0^{2\pi} x_k p(Y(\omega_k)|x_k,\theta_k)\, p(x_k,\theta_k) d\theta_k dx_k}{\int_0^\infty \int_0^{2\pi} p(Y(\omega_k)|x_k,\theta_k)\, p(x_k,\theta_k) d\theta_k dx_k}$$

Where, $\theta_k = \theta_x(k)$. Under the assumed Gaussian model $p(Y(\omega_k)|x_k,\theta_k)$ and $p(x_k,\theta_k)$ are given by

$$p(Y(\omega_k)|x_k,\theta_k) = \frac{1}{\pi\lambda_d(k)} exp\left\{-\frac{1}{\lambda_d(k)}\left|Y_k - X_k e^{j\theta_x(k)}\right|^2\right\} \tag{4}$$

$$p(x_k,\theta_k) = \frac{x_k}{\pi\lambda_d(k)} exp\left\{-\frac{X_k^2}{\lambda_x(k)}\right\} \tag{5}$$

Where, $\lambda_x(k) \triangleq E\{|X_k|^2\}$, and $\lambda_d(k) \triangleq E\{|D_k|^2\}$ are the variances of the kth spectral component of the speech and noise respectively. Substituting Eq. 4 and Eq. 5 into Eq. 3 gives

$$\hat{X}_k = \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} exp\left(-\frac{v_k}{2}\right)\left[(1+v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] Y_k \tag{6}$$

Where $\Gamma(\cdot)$ denotes the gamma function, with $\Gamma(\cdot) = \sqrt{\pi}/2$, and $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively. The variable, v_k is defined by

$$v_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k \tag{7}$$

Where ξ_k and γ_k are interpreted as the a priori and a posteriori signal-to-noise ratio (SNR), respectively and are defined by

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \tag{8}$$

$$\gamma_k \triangleq \frac{Y_k^2}{\lambda_d(k)} \tag{9}$$

At high SNR, $\xi_k \gg 1$ and $\gamma_k \gg 1$; therefore, the estimator can be simplified as:

$$\hat{X}_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k \tag{10}$$

The above is called as Wiener estimator. Because of its several advantages, the MMSE estimation of speech spectrum have received considerable attention; however, the existing related methods have been reported several limitations either on the underlying assumptions or derivation of the estimators. Therefore, a Laplacian based MMSE estimator is taken in this paper.

The basic idea of Laplacian based MMSE STSA estimator is to find the optimal estimate of the modulus of speech signal DFT components. It is based on the assumption that the real and imaginary parts of these components are modeled by a Laplacian distribution. The noise signal DFT components are assumed to be Gaussian distributed. The Laplacian estimator has been discussed before; however, it is presented here to determine the speech-presence uncertainty, required to estimate source location. It is because in a typical speech signal, it is very likely that speech is not present at all times. It is also because running speech contains a great deal of pauses, even during speech activity. The stop closures, for example, which are brief silence periods occurring before the burst of stop consonants, often appear in the middle of a sentence. Also, speech might not be present at a particular frequency even during voiced speech segments. Therefore, a two-state model for speech events is considered, which is based on the fact that either speech is present at a particular frequency bin (hypothesis H_1) or that is not (hypothesis H_0). This is expressed mathematically using the following binary hypothesis model:

$$H_0^k : speech\ absence : Y(\omega_k) \quad (11)$$

$$H_1^k : speech\ present : Y(\omega_k) = X(\omega_k) + D(\omega_k) \quad (12)$$

To incorporate the above binary model to an MMSE estimator, a weighted average of two estimators is used. So, if the original MMSE estimator had the form $\hat{X}_k = E(X_k|Y(\omega_k))$, then the new estimator, has the form:

$$\hat{X}_k = E(X_k|Y(\omega_k), H_1^k)P(H_1^k|Y(\omega_k)) + E(X_k|Y(\omega_k), H_0^k)P(H_0^k|Y(\omega_k)) \quad (13)$$

Where $P(H_1^k|Y(\omega_k))$ denotes the conditional probability that speech is present in frequency bin *k*, given the noisy speech spectrum. Similarly $P(H_0^k|Y(\omega_k))$ denotes the conditional probability that speech is absent given the noisy speech spectrum. The term $E(X_k|Y(\omega_k), H_0^k)$ in the above equation is zero since it represents the average value of X_k given the noisy spectrum $Y(\omega_k)$ and the fact that speech is absent. Therefore, the MMSE estimator mentioned above reduces to

$$\hat{X}_k = E(X_k|Y(\omega_k), H_1^k)P(H_1^k|Y(\omega_k)) \quad (14)$$

The term $P(H_1^k|Y(\omega_k))$ can be computed using Bayes rule. The MMSE estimator of the spectral component at frequency bin k is weighted by the probability that speech is present at that frequency:

$$P(H_1^k|Y(\omega_k)) = \frac{p\left(Y(\omega_k)\middle|H_1^k\right)P(H_1)}{p\left(Y(\omega_k)\middle|H_1^k\right)P(H_1^k) + p\left(Y(\omega_k)\middle|H_0^k\right)P(H_0)} = \frac{\Lambda(Y(\omega_k)|q_k)}{1 + \Lambda(Y(\omega_k)|q_k)} \quad (15)$$

Where $\Lambda(Y(\omega_k), q_k)$ is the generalized likelihood ratio defined by:

$$\Lambda(Y(\omega_k), q_k) = \frac{1 - q_k}{q_k} \frac{p\left(Y(\omega_k)\middle|H_1^k\right)}{p\left(Y(\omega_k)\middle|H_0^k\right)} \quad (16)$$

where $q_k = P(H_0^k)$ denotes the a priori probability of speech absence for frequency bin k. The a priori probability of speech presence i.e. $P(H_1^k)$ is given by $1 - q_k$. Theoretically, the optimal estimate under hypothesis H_0^k is identical to zero but a small nonzero value might be preferable for perceptual purposes. Under hypothesis H_0 , $Y(\omega_k) = D(\omega_k)$, and given that the noise is complex Gaussian with zero mean and variance $\lambda_d(k)$; it follows that $p\left(Y(\omega_k)\middle|H_0^k\right)$ will also have a Gaussian distribution with the same variance, i.e.,

$$p\left(Y(\omega_k)\middle|H_0^k\right) = \frac{1}{\pi\lambda_d(k)} \exp\left(-\frac{Y_k^2}{\lambda_d(k)}\right) \quad (17)$$

If $X(\omega_k)$follows a Laplacian distribution, it is required to compute $p\left(Y(\omega_k)\middle|H_1^k\right)$. Assuming independence between real and imaginary components, we have:

$$p\left(Y(\omega_k)\middle|H_1^k\right) = p(z_r, z_i) = p_{z_r(k)}(z_r)p_{z_i(k)}(z_i) \quad (18)$$

where $z_r(k) = Re\{Y(\omega_k)\}$ and $z_i(k) = Re\{Y(\omega_k)\}$. Under hypothesis H_1, the pdf of $Y(\omega_k) = X(\omega_k) + D(\omega_k)$ needs to be derived, where $X(\omega_k) = X_r(\omega_k) + jX_i(\omega_k)$ and $D(\omega_k) = D_r(\omega_k) + jD_i(\omega_k)$. The pdfs of $X_r(\omega_k)$ and $X_i(\omega_k)$ are assumed to be Laplacian and the pdfs of $D_r(\omega_k)$ and $D_i(\omega_k)$ are assumed to be Gaussian with variance $\sigma_d^2/2$ and zero mean. Its implementation and results obtained are discussed in section III.

B. Step 2: Locating the speech source

The research reported in this paper is limited to locating sound source relative to only one microphone in two dimensional space [10-12]. It is assumed that the source is moving in the line of sight plane within the two microphone positions. The time delay and location is estimated relative to one microphone as shown in Figure 1.

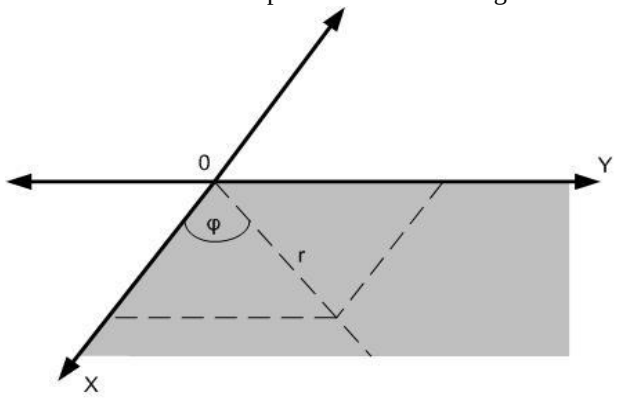

Figure 1: Microphone array based speech source locator.

If the speech source is located at *r* distance from a microphone and at the azimuth angle φ, then its location relative to one microphone is given by:

$$r = [x, y] \quad (19)$$

And, its direction related to one microphone is given by direction parameter *k*, as given by:

$$k = \begin{bmatrix} k_x \\ k_y \end{bmatrix} = \begin{bmatrix} \cos(\varphi) \\ \sin(\varphi) \end{bmatrix} \quad (20)$$

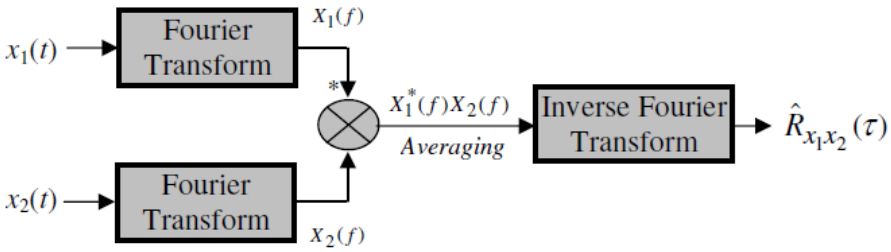

Figure 2: Basic principle of cross-correlation for time delay estimation of speech signals in microphone arrays.

Time delay estimation of sound signals arrived at microphone arrays is based on the principle of cross-correlation of sound signals (Figure 2). A distinct peak is obtained in the correlation peaks, whenever the speaker speaks and the same signals arrived at microphone arrays. If the time taken by these speech signals to travel from speech source position to microphone positions 1 and 2 be t_1 and t_2 respectively, then the time delay (τ) between two measured signals ($x_1(t)$ and $x_2(t)$) is given by

$$\tau = -\frac{r_2 - r_1}{c} \quad (21)$$

Where, c is the speed of sound, r_2 and r_1 are the positions of speech source relative to two microphones (Figure 1). If multiple microphones are used in various directions then the time delay of signals at each microphone relative to one microphone is given as:

$$\tau = [\tau_{12}, \tau_{13}, \cdots \tau_{1j}] \quad (22)$$

Where, τ_{1j} denoted the time delay of the sound propagation between the j^{th} sensor and 1^{st} sensor. For multiple microphones based array, the time delay is related to source location as

$$\tau = -\frac{R.k}{c} \quad (23)$$

Where, *R* denotes distances between each sensor and the reference sensor, and is given by:

$$R = \begin{bmatrix} r_2 - r_1 \\ \vdots \\ r_M - r_1 \end{bmatrix} \quad (24)$$

The direction parameter *k* can be calculated from time delay estimation $\hat{\tau}$ and distance as:

$$\hat{k} = -c(R^T R)^{-1} R^T \hat{\tau} \quad (25)$$

As in this paper, only the speech source in a two dimensional plane is taken, so the location is given as:

$$\varphi = tan^{-1}(\hat{k}_y / \hat{k}_x) \quad (26)$$

Time delay (τ) is estimated using crosscorrelation of measured speech signals as shown in Figure 2. The quality of this estimate depends upon the type, positioning of speech source and the processing of signals obtained. Cross-correlation of two measured speech signals $x_1(t)$ and $x_2(t)$ is usually done in the frequency domain for ease of calculations, by taking the inverse Fourier Transform (Figure 2) of the product of complex conjugate of Fourier transform of measured speech signal $x_1(t)$ with the Fourier Transform of speech signal $x_2(t)$ as

$$Rx_1x_2(\tau) = \frac{1}{T}\int_{-\infty}^{\infty} X_1^*(f)X_2(f)e^{j2\pi f\tau}df \quad (27)$$

Where, * denotes the complex conjugate. In practise, due to the noise and attenuation of speech signals from source to microphones, the measured signals and their correlation function fluctuates. These fluctuations can sometimes be very large, making it difficult to identify the correlation peak responsible for speech source signals. Therefore, the cross-correlation function $\rho x_1x_2(\tau)$ is expressed in normalised (dimensionless) form i.e. on the scale of -1 to 1, as:

$$\rho x_1x_2(\tau) = \frac{\hat{R}x_1x_2(\tau)}{\sqrt{\hat{R}x_1x_1(0)\hat{R}x_2x_2(0)}} \quad (28)$$

Where, $\hat{R}x_1x_1(0)$ and $\hat{R}x_2x_2(0)$ are the values of auto-correlation functions $\hat{R}x_1x_2(\tau)$ and $\hat{R}x_1x_2(\tau)$ at τ=0. Its experimental validation is discussed below.

III. RESULTS & DISCUSSION

In order to evaluate the performance of proposed MMSE based estimator and speech source localization, the same is implemented in Matlab. A small length of approximately 0.5 seconds of speech of a male person and noise (moving fan) is recorded using microphone based array (Figure 3). The recorded speech signals are shown in Figures 4 and 5. These noisy speech signals are down-sampled to 8KHz. The proposed estimators were applied to 20ms duration frames of speech using a Hamming window with 50% overlap between frames. The enhanced signal was combined using the overlap and add approach. The STSA of the speech signals was then estimated and combined with the complex exponential of the noisy phase. The proposed decision–directed approach is used in the proposed estimators to compute the a priori SNR ξ_k, with $\alpha = 0.98$. The a priori probability of speech absence, q_k, is set to $q_k = 0.3$.

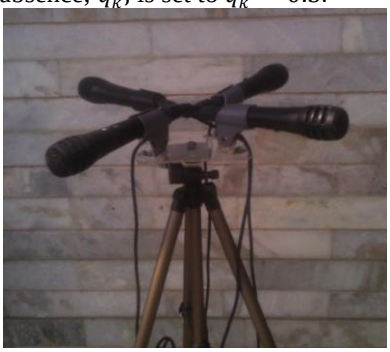

Figure 3: The designed microphone array.

Figures 4 and 5 show the speech signal enhanced by the proposed MMSE estimator and its corresponding correlation results for distance location. It is clear that the proposed estimator yielded less residual noise. It is also clear that the correlation results yield a sharp peak in situations of clean speech signals which will yield better estimation of the time delay. Thus, the proposed algorithm is capable of estimating speech signal and thus speaker location even in situations of very high noise or low SNRs can be estimated. Therefore, it can be concluded that the proposed algorithm is capable of estimating the location of speaker even in the presence of a high degree of noise.

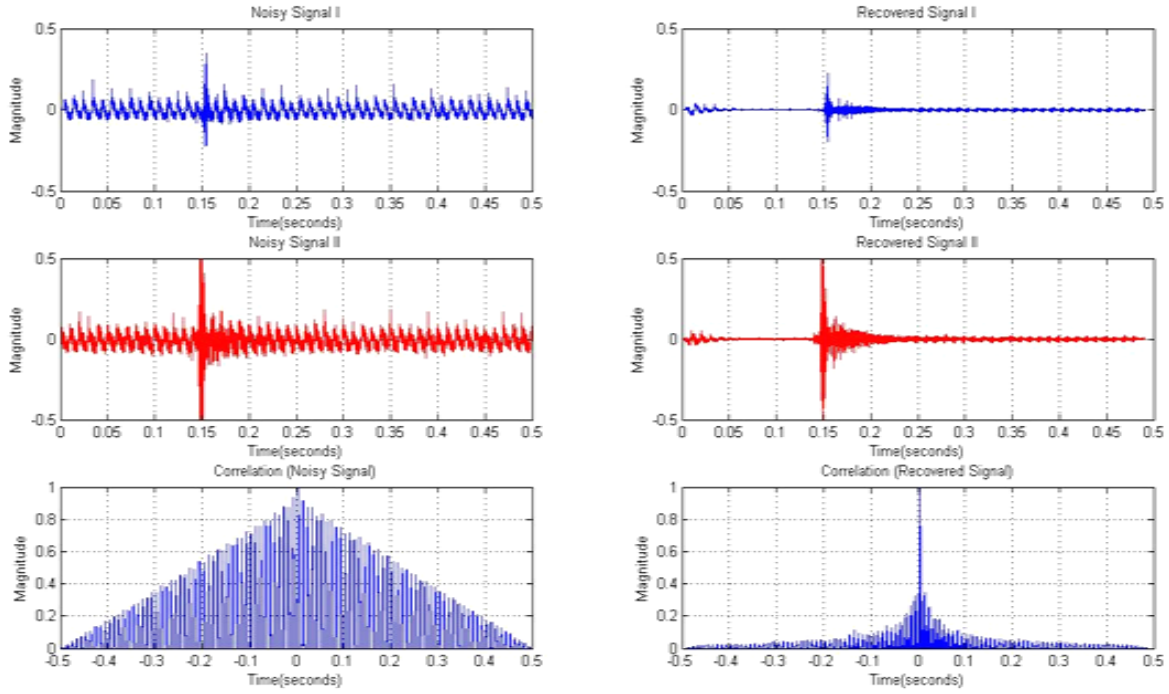

Figure 4: Example 1: Correlation of the recorded speech signals and the signals cleaned with MMSE algorithm.

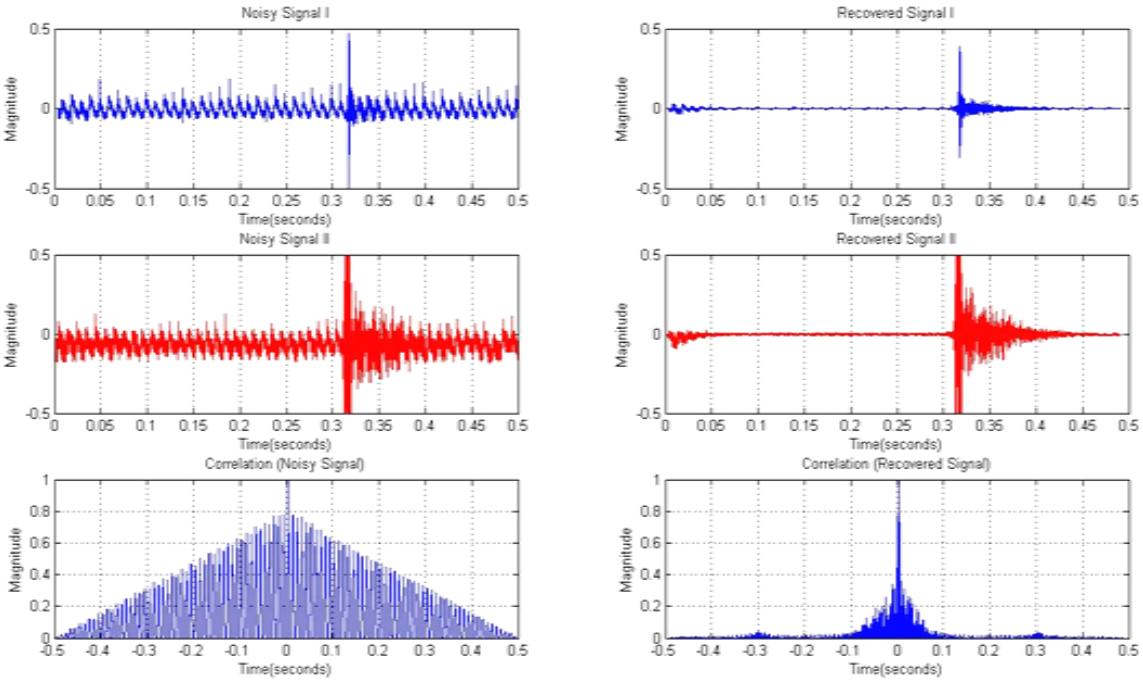

Figure 5: Example 1: Correlation of the recorded speech signals and the signals cleaned with MMSE algorithm.

IV. CONCLUSION

Because of the high level of noise (especially due to traffic), now-a-days, it is difficult to record a clean speech signal. Depending upon the SNR, the complete signal can be fade-in by noise. Therefore, a MMSE estimator based speech enhancement algorithm is proposed and evaluated in this paper. The performance of this algorithm is evaluated in terms of various analytical parameters. From the results obtained it is concluded that the proposed algorithm is capable of estimating speech signal even in situations of very high noise or low SNRs. These estimated speech signals are used to estimate the location of their source. This is done by putting these signals into a microphone array, with its output to sound card data acquisition system. The time delay between signals is estimated using cross-correlation. This is used to locate the position of the sound source.

Reference

[1] R.C. Nongpiur, "Impulse Noise Removal In Speech Using Wavelets" ICASSP, IEEE, 2008.

[2] Meng Joo Er., "Adaptive Noise Cancellation Using Enhanced Dynamic Fuzzy Neural Networks", IEEE Trans. Fuzzy Systems, vol. 13, No. 3, June 2005, pp 331-342.

[3] C. Plapous, C. Marro and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement". IEEE Trans. Audio Speech Lang. Process., pp. 2098-2108, 2006.

[4] C. Plapous, C. Marro and P. Scalart, "A two-step noise reduction technique," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, Montréal, QC, Canada, May 2004, vol. 1, pp. 289–292.

[5] Z. Chen, "Simulation of Spectral Subtraction Based Noise Reduction Method", International Journal of Advanced Computer Science and Applications, vol. 2, No. 8, 2011.

[6] M. Hasan, S. Salahuddin and M. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," IEEE Signal Processing, vol. 11, No. 4, pp. 450–453, Apr. 2004.

[7] C. F. Juang, and C. T. Lin, "Noisy Speech Processing by Recurrently Adaptive Fuzzy Filters", IEEE Transactions on Fuzzy Systems, vol. 9, No. 1, February 2001, pp 139-152.

[8] Chen, J., L. Shue, and W. Ser, "A new approach for speaker tracking in reverberant environment. Signal Processing", 2002. **82**(7): p. 1023-1028.

[9] Castellini, P. and A. Sassaroli, "Acoustic source localization in a reverberant environment by average beamforming" Mechanical Systems and Signal Processing, 2010. **24**(3): p. 796-808.

[10] Albarbar, A., Gennish, R.; Gu, F.; Ball, A. "Lubricating oil condition monitoring using vibration and air-borne acoustic measurements", Proceedings of the 7th Biennial Conference on Engineering Systems Design and Analysis - 2004, v 3, Proceedings of the 7th Biennial Conference on Engineering Systems Design and Analysis - 2004, 2004, p 205-209.

V. REFERENCES

[11] Krim, H. and M. Viberg (JULY 1996). "Two Decades of Array Signal Processing Research," IEEE SIGNAL PROCESSING MAGAZINE: 67-94.

[12] Knapp, C. and G. Carter, "The generalized correlation method for estimation of time delay," Acoustics, Speech and Signal Processing, IEEE Transactions on, 1976. 24(4): p. 320-327.

